Molecular Epidemiology and Biomarkers in Etiologic Cancer Research: The New in Light of the Old

Paolo Vineis1,2 and Frederica Perera3

1Department of Epidemiology and Public Health, Imperial College London, London, United Kingdom; 2Institute for Scientific Interchange Foundation, Torino, Italy; and 3Department of Environmental Sciences and Columbia Center for Children's Environmental Health, Mailman School of Public Health, Columbia University, New York, New York

Abstract

The purpose of this review is to evaluate progress in molecular epidemiology over the past 24 years in cancer etiology and prevention to draw lessons for future research incorporating the new generation of biomarkers. Molecular epidemiology was introduced in the study of cancer in the early 1980s, with the expectation that it would help overcome some major limitations of epidemiology and facilitate cancer prevention. The expectation was that biomarkers would improve exposure assessment, document early changes preceding disease, and identify subgroups in the population with greater susceptibility to cancer, thereby increasing the ability of epidemiologic studies to identify causes and elucidate mechanisms in carcinogenesis. The first generation of biomarkers has indeed contributed to our understanding of risk and susceptibility related largely to genotoxic carcinogens. Consequently, interventions and policy changes have been mounted to reduce risk from several important environmental carcinogens. Several new and promising biomarkers are now becoming available for epidemiologic studies, thanks to the development of high-throughput technologies and theoretical advances in biology. These include toxicogenomics, alterations in gene methylation and gene expression, proteomics, and metabolomics, which allow large-scale studies, including discovery-oriented as well as hypothesis-testing investigations. However, most of these newer biomarkers have not been adequately validated, and their role in the causal paradigm is not clear. There is a need for their systematic validation using principles and criteria established over the past several decades in molecular cancer epidemiology.

Consequently, interventions and policy changes have been mounted to reduce risk from several important environmental carcinogens. Several new and promising biomarkers are now becoming available for epidemiologic studies, thanks to the development of high-throughput technologies and theoretical advances in biology. These include toxicogenomics, alterations in gene methylation and gene expression, proteomics, and metabolomics, which allow large-scale studies, including discovery-oriented as well as hypothesis-testing investigations. However, most of these newer biomarkers have not been adequately validated, and their role in the causal paradigm is not clear. There is a need for their systematic validation using principles and criteria established over the past several decades in molecular cancer epidemiology. (Cancer Epidemiol Biomarkers Prev 2007;16(10):1954–65)

Introduction

In 1982, Perera and Weinstein (1) proposed "molecular cancer epidemiology" as a new paradigm for cancer research that incorporated biomarkers into epidemiologic studies to reveal mechanisms and events occurring along the theoretical continuum between exposure and disease. The biomarkers were categorized as markers of internal dose, markers of biologically effective dose, markers of early response/effect, or markers of susceptibility. The U.S. National Academy of Sciences in 1987 convened a workshop on the use of biomarkers in environmental health research that adopted this paradigm and expanded it to include a fourth category: altered structure and function (Fig. 1; refs. 2-5).

Consequently, interventions and policy changes have been mounted to reduce risk from several important environmental carcinogens. Several new and promising biomarkers are now becoming available for epidemiologic studies, thanks to the development of high-throughput technologies and theoretical advances in biology. These include toxicogenomics, alterations in gene methylation and gene expression, proteomics, and metabolomics, which allow large-scale studies, including discovery-oriented as well as hypothesis-testing investigations. However, most of these newer biomarkers have not been adequately validated, and their role in the causal paradigm is not clear. There is a need for their systematic validation using principles and criteria established over the past several decades in molecular cancer epidemiology. (Cancer Epidemiol Biomarkers Prev 2007;16(10):1954–65)

The original model—that of a continuum of molecular/genetic alterations leading to cancer that can be accessed using biomarkers—remains fundamentally valid. Research in molecular epidemiology in the last 20 years has followed this general model and many key research findings are supportive (6). As summarized in Table 1, evidence supporting the paradigm of cancer as a continuum involving measurable molecular/genetic events includes: (a) studies showing a correlation between external measurements of exposure and biomarkers of biologically effective dose or early biological response/effect, such as carcinogen-DNA or carcinogen-hemoglobin adducts, in relation to exposure to polycyclic aromatic hydrocarbons (PAHs; ref. 7), acrylamide (8), styrene (9), or 1,3-butadiene (9); (b) studies showing overall correlations between DNA or protein adduct levels and environmental exposures to carcinogens via smoking, the workplace, or the ambient air, with significant interindividual variation in adduct levels (10); (c) studies confirming the ability of certain carcinogen-DNA adducts (11-13) and chromosome aberrations (14) to predict cancer; and (d) studies confirming the role of certain genetic variants [single nucleotide polymorphisms (SNPs)] in modulating the risk of cancer, particularly in subjects exposed to carcinogens (15).

This review provides the platform for an assessment of the potential of new epigenetic biomarkers and "omic" technologies in cancer prevention and recommendations

Received 5/21/07; revised 7/13/07; accepted 7/23/07.

Grant support: Compagnia di San Paolo to the Institute for Scientific Interchange Foundation; European Union for the project Environmental Cancer Risk, Nutrition and Individual Susceptibility (P. Vineis); NIH; National Institutes of Environmental Health Sciences grants SPO1ES04600, 5RO1ES08977, 1R01CA127332, and R01CA69094; U.S. Environmental Protection Agency grants R827027 and RD-832141; Bauman Family Foundation; and New York Community Trust.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Requests for reprints: Frederica P. Perera, Department of Environmental Sciences and Columbia Center for Children's Environmental Health, Mailman School of Public Health, Columbia University, 100 Haven Avenue, Tower 3, #22F, New York, NY 10032. Phone: 212-304-7280; Fax: 212-544-1943. E-mail: fpp1@columbia.edu

Copyright © 2007 American Association for Cancer Research.

doi:10.1158/1055-9965.EPI-07-0457

Cancer Epidemiol Biomarkers Prev 2007;16(10). October 2007
Figure 1. Updated model for molecular epidemiology. Adapted from Perera and Weinstein (1), National Research Council (2), Schulte et al. (3), and Harris (4).

for next steps in molecular epidemiologic research. In the next years, the use of new technologies/markers, such as proteomics, metabolomics, and epigenomics, will become highly relevant, particularly for longitudinal epidemiologic longitudinal studies of chronic disease. When combined with the best of the earlier validated biomarkers of dose, effect, and susceptibility, such new markers have the potential to add considerably to knowledge about the mechanistic pathways that relate pathogenic exposures to disease onset and also to serve as informative early markers of disease risk. However, research is needed to establish the validity and applicability of these new biomarkers/technologies.

In the following paragraphs, we consider the lessons we can draw from the three related areas of biomarker use: etiologic discovery, carcinogen identification and evaluation, and primary prevention. Rather than an encyclopedic review, we have chosen to present several paradigmatic examples of each area. Among promising biomarkers and technologies not included here are those related to inflammation and obesity (15, 16), genomewide scans (16), and tumor markers (17).

The Use of Biomarkers in Etiologic Cancer Research: What Have We Learned? The Examples of Benzene and Tobacco Smoke/PAHS

Benzene. The example of benzene and hematologic malignancies is paradigmatic because it combines markers of several different types that belong to the carcinogenesis pathway shown in Fig. 1. Some of the most interesting results have come from investigations using early preclinical response markers. Prospective

### Table 1. Discoveries that support the original model of molecular epidemiology

<table>
<thead>
<tr>
<th>Marker linked to exposure or disease</th>
<th>Exposure</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal dose</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urinary metabolites (NNK, NNN)</td>
<td>Nitrosocompounds in tobacco</td>
<td>Hecht et al. (41, 89)</td>
</tr>
<tr>
<td>Biologically effective dose</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNA adducts</td>
<td>PAHs, aromatic compounds</td>
<td>Tang et al. (11)</td>
</tr>
<tr>
<td>Albumin adducts</td>
<td>AFB1</td>
<td>Groopman et al. (90)</td>
</tr>
<tr>
<td>Hemoglobin adducts</td>
<td>Acrylamide</td>
<td>Santella et al. (91)</td>
</tr>
<tr>
<td></td>
<td>Styrene</td>
<td>Hatma et al. (8)</td>
</tr>
<tr>
<td></td>
<td>1,3-Butadiene</td>
<td>Vodicka et al. (9)</td>
</tr>
<tr>
<td>Preclinical effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chromosome aberrations</td>
<td>Exposure and/or cancer</td>
<td>Bonassi et al. (14)</td>
</tr>
<tr>
<td></td>
<td>Lung</td>
<td>Smith et al. (85)</td>
</tr>
<tr>
<td></td>
<td>Leukemia</td>
<td>Holecová et al. (93)</td>
</tr>
<tr>
<td></td>
<td>Benzene</td>
<td>Perera et al. (94)</td>
</tr>
<tr>
<td>HPRT</td>
<td>PAHs</td>
<td>Ammenheuser et al. (95)</td>
</tr>
<tr>
<td></td>
<td>1,3-Butadiene</td>
<td>Lee et al. (96)</td>
</tr>
<tr>
<td>Glycophorin A</td>
<td>PAHs</td>
<td>Gwosdz et al. (97)</td>
</tr>
<tr>
<td>Gene expression</td>
<td>Cisplatin</td>
<td>Berwick and Vineis (98);</td>
</tr>
<tr>
<td>Genetic susceptibility</td>
<td></td>
<td>Cheng et al. (99)</td>
</tr>
<tr>
<td>Phenotypic markers</td>
<td>e.g., DNA repair capacity in head and neck cancer</td>
<td>Berwick and Vineis (98);</td>
</tr>
<tr>
<td>SNPs</td>
<td>Bladder</td>
<td>Garcia-Closas (100);</td>
</tr>
<tr>
<td>NAT2, GSTM</td>
<td>Lung</td>
<td>Cheng et al. (99)</td>
</tr>
<tr>
<td>CYP1A1</td>
<td></td>
<td>Vineis et al. (101)</td>
</tr>
</tbody>
</table>

Abbreviations: NNK, 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone; NNN, N-nitrosornornicotine; HPRT, hypoxanthine phosphoribosyltransferase.
studies have shown that, at the population level, chromosome aberrations are able to predict the onset of cancer, including hematologic malignancies. Combined analyses of data from Nordic and Italian prospective cohort studies involving 3,541 subjects found that chromosomal aberrations were significant predictors of cancer (14). In the Italian cohort, cancer predictivity of high chromosomal aberrations was greatest for hematologic malignancies, with a standardized mortality ratio of 5.49 (95% confidence interval, 1.49-140.5; ref. 18).

Specific chromosomal aberrations have been observed in both preleukemia and leukemia patients exposed to benzene as well as in otherwise healthy benzene-exposed workers (19). By use of fluorescent in situ hybridization and PCR, Zhang et al. found that high occupational benzene exposure increased the frequencies of aberrations in chromosomes 5, 7, 9, 8, and 11, aberrations that are frequently seen in acute myeloid leukemias and in preleukemic myelodysplastic syndrome.

In the same studies on Chinese workers, protein expression patterns were detected by surface-enhanced laser desorption/ionization-time of flight, a technique widely used in proteomic research (see below). Surface-enhanced laser desorption/ionization-time-of-flight analysis of exposed and unexposed subjects revealed that lower expression of PF4 and CTAP-III proteins is a potential biomarker of the early biological effects of benzene and may play a role in the immunosuppressive effects of benzene (20).

Finally, Lan et al. (21) investigated 20 candidate susceptibility genes in the same Chinese cohort. After accounting for multiple comparisons, SNPs in five genes were associated with a statistically significant decrease in total WBC counts among exposed workers [IL-1A (−889C>T), IL-4 (−1098T>G), IL-10 (−819T>C), IL-12A (8685G>A), and VCAM1 (−1591T>C)]. This report provides evidence that SNPs in genes that regulate hematopoiesis influence benzene-induced hematotoxicity. (We note that an association between certain SNPs and cancer in one study does not necessarily translate to other populations and exposure scenarios and must be confirmed).

**Tobacco Smoke/PAHs and Lung Cancer.** Most of the molecular epidemiologic research on lung cancer has targeted tobacco smoke and PAHs as model carcinogens. PAHs are found in tobacco smoke, in outdoor air from fossil fuel combustion, in indoor air from cooking, heating, and smoking, and in the diet (22). PAHs are animal carcinogens, including transplacentally, and human carcinogens (23, 24). PAHs, such as benzo(a)pyrene, form adducts with DNA, a mechanism considered to be a carcinogens (23, 24). PAHs, such as benzo(a)pyrene (32, 33); and (c) similar to the example of benzene, certain SNPs or genes involved in the metabolism or detoxification of PAHs or in the repair of PAH DNA adducts have been implicated as effect modifiers in lung carcinogenesis.

More recently, as will be discussed in “Categories of Epigenetic and Omic Biomarkers,” epigenetic mechanisms have emerged as important in lung cancer related to tobacco smoking (34-36).

**Conclusion.** In summary, studies using biomarkers of biologically effective dose (adducts in the case of tobacco smoke/PAHs) or preclinical effect (chromosomal aberrations and proteomics in the case of benzene; p53 for tobacco smoke/PAHs) and individual susceptibility (SNPs) have been valuable in elucidating the steps that link benzene to the onset of leukemia and tobacco smoke/PAHs to the onset of lung cancer.

**The Use of Biomarkers/Molecular Epidemiology in Cancer Risk Assessment and Prevention**

**Biomarkers in Hazard Evaluation: The Example of the IARC Monographs.** Biomonitoring and molecular epidemiologic studies have provided mechanistic data on carcinogens that have been used in risk assessment and in some cases regulation of those carcinogens. The Monographs on the Evaluation of Carcinogenic Risks to Humans have been published by the IARC since 1971 as a guide to regulatory and public health agencies in their decision making. The Monographs report the assessment made by groups of experts of the weight of evidence in humans and experimental animals based on detailed guidelines (see the IARC Web site). The evidence in humans or animals is graded using the categories of “sufficient,” “limited,” “inadequate,” and “evidence of lack of carcinogenicity.” Then, animal and human evidence is combined for the overall evaluation that consists of classifying the agents into group 1 (carcinogenic to humans), group 2A (probably carcinogenic to humans), group 2B (possibly carcinogenic to humans), group 3 (not classifiable), and group 4 (probably not carcinogenic to humans). Since 1997 (Monograph 54), mechanistic evidence, including biomarker data in humans or animals, has been used to “upgrade” or “downgrade” the classification of the agents.

Table 2 shows the chemicals or exposures for which mechanistic data contributed substantially to the final evaluation of carcinogenicity and Appendix A explains the underlying rationale. The application of mechanistic...
Evidence from biomarkers and molecular epidemiology to carcinogen hazard identification is justified by the fact that many animal carcinogens have not been adequately studied epidemiologically. This situation is not likely to be due to chance because there is usually little motivation to fund studies in human populations exposed to specific chemicals, the statistical power of studies that are conducted tends to be low, and exposure assessment is difficult. As a consequence, these studies are few and difficult to interpret at best. Molecular evidence in humans can reasonably be integrated with the experimental data to fill the gap.

The Use of Biomarkers in Exposure Intervention, Chemoprevention, and Dietary Preventive Trials

**Biomarkers of DNA Damage in Preventive Trials.** Primary prevention encompasses a spectrum of measures that include reduction and avoidance of exposure, inhibition of carcinogen activation after it has entered the body, blocking interactions with the genome, and suppression of the propagation of premalignant changes. Many preventive trials have used DNA damage as an intermediate biomarker or end point. Examples of molecular epidemiologic studies that have documented the benefits of reduction of exposure include studies of smokers enrolled in smoking cessation programs. In one study, blood samples were drawn from 400 participants before they began the program and then at multiple time points after smoking cessation. Forty of the 400 smokers were successful at quitting (compliance was verified by assaying cotinine levels in blood). Levels of the biomarkers, PAH-DNA, and 4-aminobiphenyl-hemoglobin adducts were significantly reduced by 8 weeks after quitting, reflecting cessation (37). Similarly, following a reduction in air concentrations of PAHs in a Finnish iron foundry, both PAH-DNA and aromatic-DNA adduct levels in workers’ blood samples declined significantly (8). In a follow-up to the smokers’ cessation study, a randomized clinical trial of vitamins E and C in smokers found that, among treated women, but not treated men, this was a significant decline in PAH-DNA adducts compared with controls, suggesting a benefit of antioxidants in reducing procarcinogenic DNA damage in female smokers and that hormonal factors may be important in modulating antioxidant effects.

### Table 2. Chemicals or agents for which evidence from mechanistically relevant biomarkers was used in the assessment of carcinogenic hazards to humans within the IARC Monographs program

<table>
<thead>
<tr>
<th>Mechanistic evidence used to upgrade hazards from 2A (probably carcinogenic to humans) to 1 (carcinogenic to humans)</th>
<th>Ethylene oxide</th>
<th>Neutrons</th>
<th>2,3,7,8-TCDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mechanistic evidence used to upgrade hazards from 2B (possibly carcinogenic to humans) to 2A (probably carcinogenic to humans)</td>
<td>Acrylamide</td>
<td>Adriamycin</td>
<td>Azacitidine</td>
</tr>
<tr>
<td>Benz(a)anthracene</td>
<td>1,2-Dimethylhydrazine</td>
<td>Benzo(a)pyrene</td>
<td>Captofol</td>
</tr>
<tr>
<td>Chloramphenicol</td>
<td>N-ethyl-N-nitrosourea</td>
<td>CCNU</td>
<td>Cisplatin</td>
</tr>
<tr>
<td>Ethylene oxide</td>
<td>4,4'-methylenebis-2-chloroaniline; 2,3,7,8-TCDD, 2,3,7,8-tetrachlorodibenzo-p-dioxin; MNNG, N'-methyl-N'-nitro-N-nitrosoguanidine; CCNU, 1-(2-chloroethyl)-3-cyclohexyl-L-nitrosourea.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethylene dibromide</td>
<td>MOCA</td>
<td>Methyl methanesulfonate</td>
<td>MNNG</td>
</tr>
<tr>
<td>Styrene-7,8-oxide</td>
<td>Teniposide</td>
<td>Tris(2,3-dibromopropyl) phosphate</td>
<td>Vinyl bromide</td>
</tr>
<tr>
<td>UV radiation A, B, C</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Abbreviations:** MOCA, 4,4'-methylenebis-2-chloroaniline; 2,3,7,8-TCDD, 2,3,7,8-tetrachlorodibenzo-p-dioxin; MNNG, N'-methyl-N'-nitro-N-nitrosoguanidine; CCNU, 1-(2-chloroethyl)-3-cyclohexyl-L-nitrosourea.
antioxidants, investigators have used markers such as 8-hydroxy-2′-deoxyguanosine and the comet assay as intermediate markers in interventions (38). Moller and Loft (39, 40) have reviewed these intervention studies and concluded that most had extremely low statistical power (sample size usually ≤20), that the interventions led to modest changes in 8-hydroxy-2′-deoxyguanosine (~10%), and that single doses of antioxidants seemed to be more effective than multiple doses, whereas the type of antioxidant was not crucial.

Other research, not reviewed here, has used urinary metabolites of tobacco-specific carcinogens as intermediate biomarkers in chemoprevention studies [see Hecht et al. (41) for example]. In conclusion, promising markers are available, such as DNA adducts, oxidative damage, and urinary metabolites, but they await application in large-scale and well-designed trials.

**Enzyme Induction as a Biomarker/Target in Preventive Trials.** A wide range of chemicals that variously interact with cellular functions has been studied experimentally in the attempt to block the processes that lead to carcinogenesis (42, 43). Examples of anti-inflammatory compounds having *in vitro* chemopreventive activity are piroxicam, sulindac, aspirin, celecoxib, and curcumin. The selective estrogen receptor modulators, tamoxifen and raloxifene, are beneficial in the prevention of estrogen-dependent tumors. Retinoids and vitamin A derivatives, such as targretin and fenretinide, have also been investigated in the prevention of tumors as have compounds containing sulfur, such as sulforaphane and oltipraz and the steroid dehydroepiandrosterone. The biomarkers used to determine the chemopreventive activity of these compounds are quite often activities of enzymes (42, 43).

Probably, the most widely studied among these compounds is oltipraz. Oltipraz has been introduced in high-risk populations based on its activity in inducing phase II enzymes (i.e., the detoxification-induction paradigm). However, results in smokers are not very encouraging. In one study, to determine if oltipraz could induce glutathione *S*-transferases and thereby reduce adduct levels of tobacco smoke constituents in the lungs and other target organs, chronic smokers were enrolled in one of three arms: 400 or 200 mg/wk oral oltipraz or placebo. There was no significant difference between treatment and placebo groups in PAH-DNA adducts in the treated subjects (41). However, these were small-scale studies and there is no evidence yet that the administration of oltipraz can be a reasonable strategy for cancer prevention as an alternative to the improvement of food quality. A simple and effective way of preventing aflatoxin contamination of food has been shown in an intervention trial in Africa (43). In this example, rather than acting through the expensive administration of chemopreventive agents in a very deprived context, the authors modified the practice of storing food, thus preventing degradation and contamination from molds producing aflatoxin.

**New Epigenetic and Omic Biomarkers**

**Categories of Epigenetic and Omic Biomarkers.** Several new and exciting biomarkers are becoming available for epidemiologic studies, thanks to the development of high-throughput technologies and theoretical advances in biology. However, most of these markers have not yet been validated, and their role in the causal paradigm is not clear. An exhaustive review is not possible here, and we refer the reader to other critical reviews, particularly for gene expression and toxicogenomics (the study of the complex interaction between the genome and chemicals in the environment or drugs, cells as they relate to disease causation; refs. 34, 35, 43, 45–47).

Here, we describe some achievements and some methodologic issues raised by the emerging fields of epigenetics and omics.

**Epigenetics and Promoter Methylation.** Epigenetic mechanisms of carcinogenesis (i.e., mechanisms that do not depend on structural changes in DNA but on functional regulation such as DNA methylation) are increasingly identified as key steps in the pathway from exposure to cancer. DNA methylation is an important epigenetic determinant of gene expression because it determines the process by which the instructions in genes are converted to mRNA, directing protein synthesis (36). DNA methylation (i.e., the covalent addition of methyl groups (CH₃) to cytosine that precedes a guanosine in the DNA sequence (the CpG dinucleotide)) occurs naturally and is thought to have a role in suppressing gene expression. CpG dinucleotides are enriched in the promoter regions of genes (CpG islands), and for this reason, methylation is thought to be involved in gene expression. Hypermethylation of promoter regions is associated with gene transcriptional silencing and is a common mechanism for the inactivation of tumor suppressor genes in human cancer (48). DNA methylation is heritable (i.e., it can pass from one generation of cells to the next).

Several genes are commonly the target of promoter hypermethylation, for example in lung cancer, including the *p16* gene (p161 NK4a/CDKN2A), DAPK, RAR-β, RASSF1, and OMGMT (a DNA repair gene; ref. 49). Global hypomethylation has also been reported (50). Both current and former smoking have been associated with aberrant *p16*, DAPK, RASSF1A, and RAR-β methylation (49). A recent investigation found that two alternative pathways could be detected in the biopsies of smoking and nonsmoking lung cancer patients, one involving methylation and *K-ras* mutations and the other *EGFR* mutations in the absence of gene methyl-
Promoter methylation in multiple genes (including those mentioned above) in the sputum was able to predict lung cancer onset with sensitivity and specificity of 64% (52). Notably, aberrant promoter methylation can be detected in the plasma of lung cancer patients (53), and a recent article by Russo et al. (54) described high frequencies of ECAD and DAPK methylation in lymphocytes of smokers versus nonsmokers. The capacity of some airborne particulate carcinogens (including those from tobacco smoke) to induce hypermethylation in the regulatory regions of tumor suppressor genes has also been investigated in animal studies (55). Overall, the animal models support involvement of promoter methylation and other epigenetic mechanisms in carcinogen-induced lung carcinogenesis (52, 56). Acetylation is another mechanism that is key in epigenetic pathways, although it has been studied less extensively than methylation in cancer epidemiology (57).

Metabonomics. Metabonomics is the study of the complete set of low-molecular weight metabolites present in a cell or organism at any time. With high-throughput techniques (nuclear magnetic resonance spectroscopy and liquid chromatography-mass spectrometry), it is possible to measure a large number of metabolites simultaneously and to define individual metabolic profiles that can be used to predict the onset of common diseases (58). Use of data processing and chemometric models has already allowed the characterization of certain disease states and metabolic disorders (58). Although several cross-sectional metabonomic studies of various cancers have been undertaken (59, 60), no longitudinal study has yet been carried out. Few validation studies of metabonomics have been published. One investigation has analyzed repeat samples from dietary studies (61). As part of a major phenotyping investigation, the authors used high-resolution 1H nuclear magnetic resonance spectroscopy to characterize 24-h urine specimens obtained from population samples in Japan (n = 259), Chicago, Illinois (n = 315), and China (n = 278). They investigated analytic reproducibility, urine specimen storage procedures, interinstrument variability, and split specimen detection. The multivariate analytic reproducibility of the nuclear magnetic resonance screening platform was >98%, and most classification errors were due to heterogeneity in handling of urine specimens. In addition, cross-population differences in urinary metabolites could be related to genetic, dietary, and gut microbial factors.

Proteomics. Proteomics is the analysis of the total protein output encoded by the genome. Proteomic research, to date, has mainly involved proteomic pattern profiling of tissue and body fluids by mass spectrometry with sophisticated bioinformatic tools to identify proteins within the complex proteomic profile that discriminate between normal, benign, or disease states (52, 56). Proteomic approaches have been used with some success for the molecular classification of tumors (62) and to develop a discriminatory pattern that distinguishes normal sera from that of ovarian, prostate, breast, lung cancer, and, more recently, cutaneous T-cell lymphoma (56, 63-70). To a lesser extent, proteomics has been applied to study physiologic or pathologic changes associated with external “environmental” exposures. Proteomic approaches have identified, for example, changes in proteins associated with oxidative stress (71). The investigation of proteomic patterns could be a powerful tool for the identification of intermediate changes that relate environmental exposures to disease onset and as an early marker of cancer. However, methodologic issues need to be resolved before application in prospective studies. In a critique of early articles, Diamandis (72) identified several methodologic problems. These included the lack of reproducibility of analytic methods; the lack of reproducibility of proteomic patterns in different series of patients and by different laboratories; unresolved effects of different protocols for sample collection, processing, and storage; possible selection effects in cases and controls (bias, confounding) partly because of the opportunistic sampling that characterized the early studies; the possible effect of drugs/other treatments; and inappropriate or nonreproducible data analysis. Many of these concerns apply to other epigenetic and omic biomarker studies as well and have been successfully addressed in more recent studies.

In conclusion, for the epigenetic and omic technologies, systematic validation studies are urgently needed.

Incorporation into Etiologic Studies of New Intermediate Biomarkers Using Omic Technologies. It is timely to discuss the challenge of incorporating epigenetic and omic biomarkers into the molecular epidemiology of cancer because, in contrast to traditional methods, high-throughput omic and epigenetic technologies allow massive investigations not based on a priori hypotheses. Such intermediate markers (either reflecting early preclinical effects or early disease) could be used for etiologic purposes (to investigate the causes and mechanisms of disease onset) or for clinical purposes (early diagnosis, prognosis, and follow-up). We will refer exclusively to the former, but many of our considerations apply to the clinical purposes as well (see refs. 73, 74 for a review of biomarker-based tools for cancer screening, diagnosis, and treatment).

As we have seen, the term intermediate marker has been used most often in the context of chemoprevention. Here, we have used the term in a very broad sense to encompass all measurable markers (in body fluids or in cells) that lie within the putative causal pathway linking an exposure to the onset of disease. If we refer to the classic scheme (Fig. 1), intermediate markers can play a role in each of the steps: they can be related to exposure (e.g., metabolomics), they can be related to early changes in the causal pathways leading to disease (such as promoter methylation, gene mutations, or telomere length), or they can express epiphenomena of preclinical disease (e.g., mutations in plasma DNA as a consequence of tumor apoptosis). It is very important that the biological significance of a marker is made explicit beforehand because false expectations can arise as a consequence of an erroneous interpretation of the role of a biomarker. For example, some markers (those on the right side of the scheme) have clinical relevance or can be useful for screening, others cannot. In the study of etiology, there are at least two important roles that intermediate markers can play: (a) to increase sensitivity and/or an earlier detectability in comparison with other markers and (b) to show biological plausibility for an exposure proposed as etiologically relevant.
Past experience with more traditional biomarkers is relevant for several reasons: it has confirmed the main theories on carcinogenesis, with evidence coming from disparate fields such as epidemiology and molecular biology; it has shown that very rarely does a single biomarker allow an exhaustive understanding of the carcinogenesis process and integration of biomarkers is necessary; and it has highlighted the many facets of biomarker validation. At the same time, the current era is different from the past and deserves special attention for the following reasons: (a) omic and new epigenetic methods tend to be discovery-oriented rather than oriented to testing specific hypotheses, (b) the main feature of current technologies is the ability to do massive testing of markers (i.e., thousands of markers at a time, potentially in thousands of subjects), and (c) such new intermediate markers introduce increased potential for confounding. So, although our ability to measure new intermediate markers has considerably increased, making the current phase potentially very exciting, methodologic challenges have expanded more than proportionally. In fact, much uncertainty surrounds the validity and applicability of new technologies [see Ransohoff (75) for example].

Another important difference between the newer and the earlier biomarkers is that the traditional cancer paradigm was very much centered around DNA and mutations, whereas recent research has uncovered several additional intermediate steps between genotype and phenotype and the importance of gene expression/modulation for carcinogenesis. Therefore, we expect that new biomarkers will be less centered around mutations, although combinations of both types of biomarkers could be informative because pathways are not mutually exclusive.

Validating Promising Intermediate Markers. The early story of proteomics is an example of the risks implicit in the premature use of a technology that has not been sufficiently validated. In particular, a concept that is often unclear is the difference between technical and field validation. Technical validation has to do with intrinsic measurement error and analytic sensitivity. Field (or epidemiologic) validation is related to how a certain marker behaves in the population, depending on biological variability within the population (3).

Biomarker validation requires several steps. A marker may be extremely powerful in increasing our understanding of the natural history and pathogenesis of a disease but

![Figure 2. Receiver operating characteristic curve analysis of total PSA (tPSA), percent free PSA (%PSA), and PSA density (PSAD). Data are shown for 307 patients with total PSA values in the range of 2 to 4 ng/mL (A), 1,282 patients with total PSA values in the range of 2 to 10 ng/mL (B), 975 patients with total PSA values in the range of 4 to 10 ng/mL (C), and 1,809 patients with total PSA values in the range of 2 to 20 ng/mL (D). Adapted from Stephan et al. (102).](Image)
may still do very poorly as a predictor for preventive or clinical purposes. One of the most important goals of validation is to characterize the ability of the marker to predict disease and, in intervention studies, reflect the modification of the natural course of disease.

Figure 2 shows one of the main summary measures of the contribution of a biomarker to the prediction of disease onset, the receiver operating characteristic curve. The figure refers to the ability of prostate-specific antigen (PSA) levels (a serum tumor marker) to predict the presence of prostate cancer. The receiver operating characteristic curve is a measure of the overall capability of the marker to predict the disease, which is a function of its sensitivity and specificity. An "area under the curve" of 0.5 indicates random association between the marker measurement and the probability of disease onset. In the figure, the maximum area under the curve is 0.77, not a particularly good performance. Unfortunately, we do not have yet similar examples of predictive ability for newer omic biomarkers. Although the receiver operating characteristic curve is established for PSA and other traditional markers, it is still unknown for markers that could be potentially much more effective. Although large trials have yet to be completed, it seems that PSA is not very useful for population screening of prostate cancer because it does not distinguish between the rare cancers that progress to frank malignancy and lead to metastases and death and the much more frequent cancers that would remain clinically apparent. It is expected that gene expression microarrays or proteomics could be more predictive than PSA, but candidate biomarkers have yet to be identified and receiver operating characteristic curves developed.

A major aim of biomarker validation is to characterize biomarker variability. The main components of biomarker variability that affect the design and interpretation of epidemiologic studies are (a) biological variability related to the subject [i.e., variability between subjects (intersubject) and within subjects (intrasubject)]; (b) variability due to measurement error, including intra-laboratory and interlaboratory variability; and (c) random error. Methodologic issues should be discussed within the context of specific biomarker categories. When we design and analyze an epidemiologic study using biomarkers, we want to minimize total intragroup variability to identify intergroup differences (e.g., between exposed and unexposed or between diseased and healthy subjects), if they exist. Total intragroup variation is the weighted sum of intersubject, intrasubject, sampling, and laboratory variation, with weights that are inversely correlated to the numbers of subjects, number of measurements per subject, and analytic replicates used in the study design, respectively. Obviously, if we do not have detailed information, we cannot adjust for intragroup variation. Therefore, in epidemiologic studies using biomarkers, it is important to collect, whenever possible, (a) repeat samples (day-to-day, month-to-month, or year-to-year variation may be relevant depending on the marker), (b) potentially relevant information on subject characteristics that may influence intersubject variation, (c) information on conditions under which samples have been collected and laboratory analyses have been conducted (batch, assay, and specific procedures). To know more about how the variability in laboratory measurements influences study design decisions, see Rundle et al. (76).

Understanding Whether an Intermediate Biomarker Belongs to the Causal Pathway. One of the main challenges with intermediate biomarkers is to understand whether they belong to the causal pathway between exposure and disease, whether they are simply a side effect of exposure or disease, or whether their measurement is confounded by some other exposure. For example, it is likely that certain mutations are genuine intermediate markers in the causal pathway, whereas others are a consequence of the disease, such as genomic instability that arises in cancer cells.

An example of the uncertain status of a marker is the association between folic acid and colorectal cancer: the inverse association observed between the two, in fact, could be due to confounding by other dietary factors and not a causal association. A way to show that folate can contribute to the disease process independently of confounding by other risk factors would be to show that the levels of folate are associated with different genotypes for genes involved in folate metabolism, such as MTHFR, and that such genotypes also predict the disease. "Mendelian randomization" has been suggested in fact as a way to overcome confounding (i.e., by exploiting the random allocation of alleles from parents to the offspring). The association between a gene variant and a disease is not subject to the confounding by behavioral or socioeconomic factors that has led to misleading findings in many conventional observational epidemiologic studies. In the case of folic acid, because the different alleles of the MTHFR gene are independent of confounding factors (e.g., dietary habits), and are assorted randomly from one generation to the next, the finding that a polymorphism in MTHFR was associated with cancer would provide indirect proof of a genuine involvement of folate in the etiology of the disease (77).

Complementary Study Designs: The "Meet in the Middle" Concept. The U.S. National Institute for Environmental Health Sciences has published a strategic plan that raises many of the points above and, in particular, stresses the importance of integrating environmental exposures within the study of the natural history of disease. In describing the role of different biomarkers and the need for their validation, the National Institute for Environmental Health Sciences document also launches a plan for the development of new markers for exposure assessment. The underlying philosophy is expressed thus: "The study of how an environmental agent affects molecular targets, cellular function, tissue function, and organism survival will need to be related up and down a continuum of biological complexity that ultimately informs us about the etiology, pathogenesis, and distribution of disease. Scientific contributions from epidemiology, toxicology, molecular and cellular biology, bioinformatics, clinical medicine, and many other fields will need to be coordinated and integrated."6

---

6 http://www.niehs.nih.gov/external/plan2006/
We propose a related approach that emphasizes the complementarity between different types of markers and study designs: the use of complementary studies to validate new biomarkers such as gene or protein expression in the absence of large-scale prospective cohort studies or where cancer is rare. An example is childhood leukemia. Although few risk factors have been confirmed (78), several prospective molecular epidemiologic studies are suggesting an etiologic role for in utero and childhood exposure to environmental pollutants (79-81). Such prospective studies have not only obtained monitoring and biomarker data on exposure but also measured biomarkers, such as the internal dose of the toxicant, carcinogen-DNA adducts, gene mutations, chromosomal aberrations, and/or gene methylation or expression. Their small sample size (<1,000 subjects in each study) does not permit assessment of childhood acute lymphocytic leukemia (ALL) cases in a retrospective (case-control) study. The association between the identified exposure and biomarker is then retrospectively evaluated within the case-control study using available questionnaire, biomarker, or other data. If confirmed as significant risk factor, the exposure can be targeted for interventions.

Figure 3. In this theoretical example, a biomarker of preclinical effect (TEL-AML1), DNA methylation, and gene or protein expression is linked to a specific environmental exposure within a prospective cohort study. The same biomarker is found to be present in a subgroup of acute lymphocytic leukemia (ALL) cases in a retrospective (case-control) study. The association between the identified exposure and biomarker is then retrospectively evaluated within the case-control study using available questionnaire, biomarker, or other data. If confirmed as significant risk factor, the exposure can be targeted for interventions.

We propose a related approach that emphasizes the complementarity between different types of markers and study designs: the use of complementary studies to validate new biomarkers such as gene or protein expression in the absence of large-scale prospective cohort studies or where cancer is rare. An example is childhood leukemia. Although few risk factors have been confirmed (78), several prospective molecular epidemiologic studies are suggesting an etiologic role for in utero and childhood exposure to environmental pollutants (79-81). Such prospective studies have not only obtained monitoring and biomarker data on exposure but also measured biomarkers, such as the internal dose of the toxicant, carcinogen-DNA adducts, gene mutations, chromosomal aberrations, and/or gene methylation or expression. Their small sample size (<1,000 subjects in each study) does not permit assessment of childhood acute lymphocytic leukemia (ALL) cases in a retrospective (case-control) study. The association between the identified exposure and biomarker is then retrospectively evaluated within the case-control study using available questionnaire, biomarker, or other data. If confirmed as significant risk factor, the exposure can be targeted for interventions.

Figure 3. In this theoretical example, a biomarker of preclinical effect (TEL-AML1), DNA methylation, and gene or protein expression is linked to a specific environmental exposure within a prospective cohort study. The same biomarker is found to be present in a subgroup of acute lymphocytic leukemia (ALL) cases in a retrospective (case-control) study. The association between the identified exposure and biomarker is then retrospectively evaluated within the case-control study using available questionnaire, biomarker, or other data. If confirmed as significant risk factor, the exposure can be targeted for interventions.

We propose a related approach that emphasizes the complementarity between different types of markers and study designs: the use of complementary studies to validate new biomarkers such as gene or protein expression in the absence of large-scale prospective cohort studies or where cancer is rare. An example is childhood leukemia. Although few risk factors have been confirmed (78), several prospective molecular epidemiologic studies are suggesting an etiologic role for in utero and childhood exposure to environmental pollutants (79-81). Such prospective studies have not only obtained monitoring and biomarker data on exposure but also measured biomarkers, such as the internal dose of the toxicant, carcinogen-DNA adducts, gene mutations, chromosomal aberrations, and/or gene methylation or expression. Their small sample size (<1,000 subjects in each study) does not permit assessment of childhood acute lymphocytic leukemia (ALL) cases in a retrospective (case-control) study. The association between the identified exposure and biomarker is then retrospectively evaluated within the case-control study using available questionnaire, biomarker, or other data. If confirmed as significant risk factor, the exposure can be targeted for interventions.

We propose a related approach that emphasizes the complementarity between different types of markers and study designs: the use of complementary studies to validate new biomarkers such as gene or protein expression in the absence of large-scale prospective cohort studies or where cancer is rare. An example is childhood leukemia. Although few risk factors have been confirmed (78), several prospective molecular epidemiologic studies are suggesting an etiologic role for in utero and childhood exposure to environmental pollutants (79-81). Such prospective studies have not only obtained monitoring and biomarker data on exposure but also measured biomarkers, such as the internal dose of the toxicant, carcinogen-DNA adducts, gene mutations, chromosomal aberrations, and/or gene methylation or expression. Their small sample size (<1,000 subjects in each study) does not permit assessment of childhood acute lymphocytic leukemia (ALL) cases in a retrospective (case-control) study. The association between the identified exposure and biomarker is then retrospectively evaluated within the case-control study using available questionnaire, biomarker, or other data. If confirmed as significant risk factor, the exposure can be targeted for interventions.
Conclusions

In conclusion, the previous generation of biomarkers has contributed greatly to our understanding of risk and susceptibility related largely to genotoxic carcinogens. Consequently, interventions and policy changes have been mounted to reduce risk from several important environmental carcinogens. More recently, developed biomarkers have considerable potential in molecular epidemiology because they reflect another equally important mechanism of carcinogenicity: epigenetic alterations that affect the expression of genes and proteins. These biomarkers can be measured by high-throughput methods, allowing large-scale studies that are discovery-oriented and that can be followed by hypothesis-testing studies. Several large collaborative studies and consortia to coordinate research efforts and enhance statistical power by increasing sample size have already been formed to study specific biomarkers and cancers [see Boffetta et al. (86) for example]. However, there is an urgent need for systematic validation of epigenetic and omic markers using the principles established during the last decades with the earlier generation of biomarkers. Once validated, the newer biomarkers can be combined with the more traditional biomarkers in hypothesis-testing studies and interventions to reduce cancer risk.

Appendix A. Use of mechanistic data in the IARC Monographs

More than 900 agents have been evaluated since the inception of the Monograph Program in 91 volumes: 108 were in group 1, 64 in group 2A, and 240 in group 2B. Table 2 shows all the agents for which upgrading or downgrading occurred based on mechanistic evidence, including molecular epidemiologic or biomarker data. Most of the category changes involved agents (36 in number) for which the evidence was sufficient in animals but inadequate in humans. These would have been classified as 2B but mechanistic evidence justified upgrading them to 2A. Three agents were upgraded from 2A to 1. An example of the latter upgrading based on human mechanistic evidence is 2,3,7,8-tetrachlorodibenzo-p-dioxin, included in group 1, in spite of limited evidence in humans. The rationale was the following: (a) 2,3,7,8-tetrachlorodibenzo-p-dioxin is a multisite carcinogen in experimental animals that has been shown by several lines of evidence to act through a mechanism involving the aryl hydrocarbon receptor, (b) this receptor is highly conserved in an evolutionary sense and functions the same way in humans as in experimental animals, and (c) tissue concentrations (internal dose of 2,3,7,8-tetrachlorodibenzo-p-dioxin) are similar in heavily exposed human populations in which an increased overall cancer risk was observed and in rats exposed to carcinogenic dosage regimens in bioassays (87).

An example of upgrading from 2B to 2A is the chemotherapeutic agent cisplatin, which was done based on observed genotoxicity in treated patients, including sister chromatid exchanges in peripheral blood lymphocytes, antigenicity against cisplatin-DNA adducts, chromosomal aberrations, and micronuclei (88). Similar effects were observed in vitro and in experimental animals. Eight agents were downgraded from 2B to 3. However, none of these evaluations involved human biomarker data.

References

24. Andersen LM, Jones AB, Miller MS, et al. Metabolism of transpla-


